23/4/18

Data Mining
Assignment - 4

Avinash Vallineni
A20406657

Problem 1.1:-

3.1.1) Let $A = \{1, 2, 3, 4\}$, $B = \{2, 3, 5, 7\}$, $C = \{2, 4, 6\}$

$$Sim(A,B) = \frac{A \cap B}{A \cup B} = \frac{2}{6} = 1/3$$

$$Sim(A,C) = \frac{A \cap C}{A \cup C} = \frac{2}{5}$$

$$Sim(B,C) = \frac{B \cap C}{B \cup C} = 1/6$$

3.2.1)

{"The", "he", "e m", "mo", "mos", "ost", "st", "te", "ef", "ff"}

or

{ The most effective, most effective way, effective ways to, way to represent, to represent documents, represent document as, document as sets, as sets for, sets for the, for the purpose }

3.3.3)

	x	$h_1(x)$	$h_2(x)$	$h_3(x)$
a)	0	1	2	2
	1	3	5	1
	2	5	2	0
	3	1	5	5
	4	3	2	4
	5	5	5	3

Document text model

h_1	h_2	h_3
1	2	2
3	5	1
5	2	0
1	5	5
3	2	4
5	5	3

S_1	S_2	S_3	S_4
0	1	0	1
0	1	0	0
1	0	0	1
0	0	1	0
0	0	1	1
1	0	0	0

Signature Matrix:-

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

b) only h_3 hash function is a true permutation

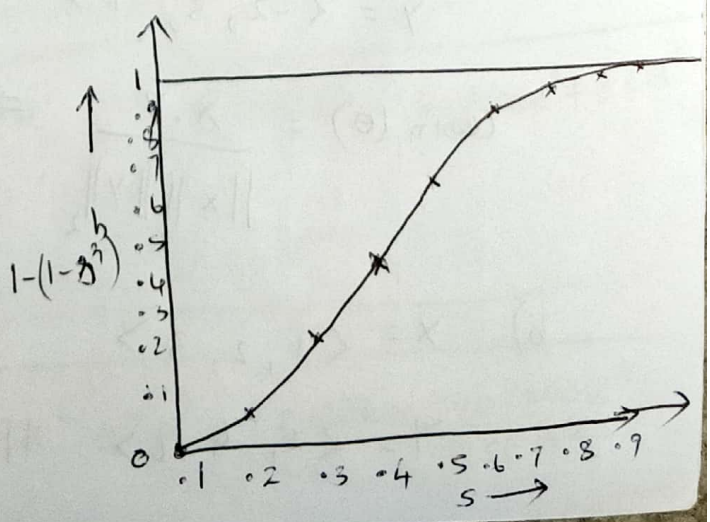
c)

Similarities	S_1-S_2	S_1-S_3	S_1-S_4	S_2-S_3	S_2-S_4	S_3-S_4
col/col	0	0	0.25	0	0.25	0.25
Sig/Sig	0.33	0.33	0.67	0.67	0.67	0.67

Jaccard similarities obtained from the Signature matrix using hash functions are not close to the similarities obtained from the Characteristic matrix.

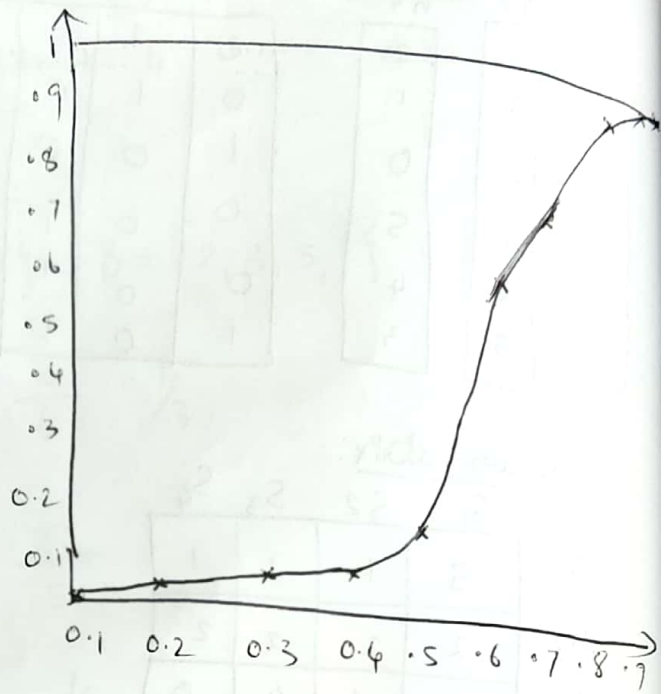
3.4.1) (i) $S: 0.1 \ldots 0.9$; $r = 3$ & $b = 10$

s	$1-(1-s^r)^b$
0.1	.01
.2	.0772
.3	.2394
.4	.4839
.5	.7369
.6	.9123
.7	.985
.8	.992
.9	.999



(1i) $r = 6$ & $b = 20$

s	$1-(1-s^r)^b$
.1	0.00002
.2	0.0013
.3	0.0145
.4	0.0788
.5	0.2702
.6	0.6154
.7	0.9182
.8	0.9977
.9	1.0



3.5.4)

a) $A = \{1, 2, 3, 4\}$ $B = \{2, 3, 4, 5\}$

$Sim(A, B) = \frac{3}{5}$

b) $A = \{1, 2, 3\}$ $B = \{4, 5, 6\}$

$Sim(A, B) = \frac{A \cap B}{A \cup B} = 0$

3.5.5)

a) $X = \langle 3, -1, 2 \rangle$ $X \cdot Y = -6 - 3 + 2 = -7$

$Y = \langle -2, 3, 1 \rangle$ $\|X\|_2 = 3.74$ $\|Y\|_2 = 3.74$

$\cos(\theta) = \frac{X \cdot Y}{\|X\|_2 \|Y\|_2} = \frac{-7}{14} = -0.5$

b) $X = \langle 1, 2, 3 \rangle$ $X \cdot Y = 2 + 8 + 18 = 28$

$Y = \langle 2, 4, 6 \rangle$ $\|X\|_2 = 3.74$ $\|Y\|_2 = 7.48$

$\cos(\theta) = \frac{X \cdot Y}{\|X\| \|Y\|} = \frac{28}{3.74 \times 7.48} = \frac{28}{28} = 1$

c) $X = \langle 5, 0, -4 \rangle$ $\quad X \cdot Y = -5 + 0 - 8 = -13$

$Y = \langle -1, -6, 2 \rangle$ $\quad \|X\| = 6.4 \quad \|Y\| = 6.4$

$\cos(\theta) = \frac{-13}{41} = -0.317$

d) $X = \langle 0, 1, 1, 0, 1, 1 \rangle$ $\quad X \cdot Y = 1$

$Y = \langle 0 0 1, 0, 0, 0 \rangle$ $\quad \|X\| = 2 \quad \|Y\| = 1$

$\cos(\theta) = \frac{1}{2 \times 1} = \frac{1}{2} = 0.5$

Problem 1.2 :-

9.21)

a)

	A	B	C
PS	3.06	2.68	2.92
DS	500	320	640
MS	6	4	6

$\cos(A,B) = \frac{A \cdot B}{\|A\| \|B\|}$

$$\cos(A,B) = \frac{8.2008 + 160000\alpha^2 + 24\beta^2}{\sqrt{9.3636 + 250000\alpha^2 + 36\beta^2}\ \sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}}$$

$$\cos(A,C) = \frac{8.9352 + 320000\alpha^2 + 36\beta^2}{\sqrt{9.3636 + 250000\alpha^2 + 36\beta^2}\ \sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

$$\cos(B,C) = \frac{7.8256 + 204800\alpha^2 + 24\beta^2}{\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}\ \sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

b)

$$\text{Cosim}(A,B)_{\alpha,\beta=1} = \frac{160032.2}{500.045 * 320.036} = \frac{160032.2}{160032.5}$$

$$= 0.9999973$$

$$\text{Cosim}(A,C)_{\alpha,\beta=1} = 0.9999953$$

$$\text{Cosim}(B,C)_{\alpha,\beta=1} = 0.9999879$$

c) $\alpha = 0.01$, $\beta = 0.5$

$$\text{Cosim}(A,B)_{\alpha=0.01,\beta=0.5} = 0.9908696$$

$$\text{Cosim}(A,C)_{\alpha=0.01,\beta=0.5} = 0.9990992$$

$$\text{Cosim}(B,C)_{\alpha=0.01,\beta=0.5} = 0.9861894$$

d) $$\alpha = \frac{1}{486.66} = .0020548$$

$$\beta = \frac{1}{5.33} = 0.1875$$

$$\text{Cosim}(A,B) = 0.9943904$$
$$\text{Cosim}(A,C) = 0.995614$$
$$\text{Cosim}(B,C) = 0.9822469$$

9.23)

A	B	C
3.06	2.68	2.92
500	320	640
6	4	6

user 1 ⇒ A → 4, B → 2, C → 5

avg = $\frac{4+2+5}{3} = 11/3$

A	B	C	row avg
4	2	5	11/3

Normalized rating

A	B	C
1/3	-5/3	4/3

b)

Processor speed = $3.06 * 1/3 - 2.68 * 5/3 + 2.92 * 4/3$

$= 1.02 - 4.466 + 3.893$

$= 0.447$

Disk size = $500 * 1/3 - 320 * 5/3 + 660 * 4/3$

$= 166.66 - 533.33 + 853.33$

$= 486.667$

Memory size = $6 * 1/3 - 4 * 5/3 + 6 * 4/3$

$= 3.333$

9.3.1

a)

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

consider as booleans

$Sim(A,B) = 4/8 = 1/2$; $Sim(A,C) = 4/8 = 1/2 = 0.5$

$Sim(B,C) = 4/8 = 1/2$

b) $Cosim(A,B) = \frac{4}{\sqrt{6} * \sqrt{6}} = 0.6667$; $Cosim(A,C) = \frac{4}{\sqrt{6} * \sqrt{6}} = 0.6667$

$Cosim(B,C) = \frac{4}{\sqrt{6} * \sqrt{6}} = 0.6667$

c)

	a	b	c	d	e	f	g	h
A	1	1	0	1	0	0	1	0
B	0	1	1	1	0	0	0	0
C	0	0	0	0	0	1	1	1

$J(A,B) = 2/5$ $\quad J_D = 1 - 2/5 = 3/5$

$J(B,C) = 1/6$ $\quad J_D = 1 - 1/6 = 5/6$

$J(A,C) = 2/6 = 1/3$ $\quad J_D = 1 - 1/3 = 2/3$

d) $Cosim(A,B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{1}{\sqrt{3}} = 0.5777$

$Cosim(B,C) = \frac{1}{\sqrt{3}\sqrt{4}} = 0.288$

$Cosim(A,C) = \frac{1}{\sqrt{4} * \sqrt{4}} = \frac{1}{4} = 0.25$

e) $avg(A) = \frac{10}{3}$ $\quad avg(C) = 3$

$avg(B) = 7/3$

	a	b	c	d	e	f	g	h
A	2/3	5/3		5/3	-7/3		-1/3	-4/3
B		2/3	5/3	2/3	-4/3	-1/3	-4/3	
C	-1		-2	0		1	2	0

$$\text{Cosin}(A,B) = \frac{52/9}{3.6515 * 2.708} = 0.5843$$

$$\text{Cosin}(A,C) = -0.11547$$

$$\text{Cosin}(B,C) = -0.739574$$

Problem 1.3 :-

5.1.1) Transition Matrix

$$\begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} 3/13, & 4/13, & 6/13 \end{bmatrix}^T$$

List:

$$\begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}, \begin{bmatrix} 0.277 \\ 0.277 \\ 0.444 \end{bmatrix}, \begin{bmatrix} 0.2314 \\ 0.3148 \\ 0.4537 \end{bmatrix}, \begin{bmatrix} 0.2345 \\ 0.3040 \\ 0.4614 \end{bmatrix},$$

$$\begin{bmatrix} 0.2301 \\ 0.3088 \\ 0.4609 \end{bmatrix} \cdots \begin{bmatrix} 0.2307 \\ 0.3076 \\ 0.4615 \end{bmatrix}$$

5.1.2)

$$V' = \beta M V + (1-\beta) e/n$$

$$= \begin{bmatrix} 4/15 & 2/5 & 0 \\ 4/15 & 0 & 2/5 \\ 4/15 & 2/5 & 2/5 \end{bmatrix} V + \begin{bmatrix} 1/15 \\ 1/15 \\ 1/15 \end{bmatrix}$$

$$v = \left[\frac{7}{27}, \frac{25}{81}, \frac{35}{81}\right]^T$$

List:-

$$\begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}, \begin{bmatrix} 0.2888 \\ 0.2888 \\ 0.4222 \end{bmatrix}, \begin{bmatrix} 0.2592 \\ 0.3125 \\ 0.4281 \end{bmatrix}, \begin{bmatrix} 0.2608 \\ 0.3070 \\ 0.4320 \end{bmatrix},$$

$$\begin{bmatrix} 0.2590 \\ 0.3090 \\ 0.4318 \end{bmatrix}, \cdots \begin{bmatrix} 0.2592 \\ 0.3086 \\ 0.4320 \end{bmatrix}$$

5.1.f)

There exist only one head node with self-direction and pageRank of 1. PageRank of the remaining nodes will be 1/2.

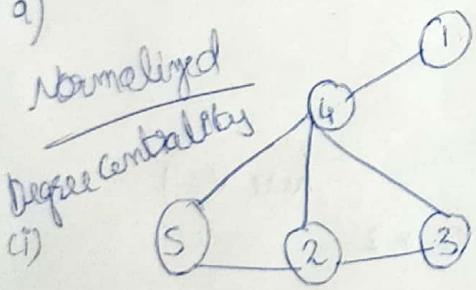
Problem 4 Centrality Measures

$C_D(v) = \deg(v)$

a)

Normalized Degree Centrality

$C_D^*(v) = \frac{1}{n-1} C_D(v)$



i)

$C_D^*(1) = 1/4 = 0.25$

$C_D^*(2) = 3/4 = 0.75$

$C_D^*(5) = 2/4 = 0.5$

$C_D^*(3) = 2/4 = 0.5$

$C_D^*(4) = 4/4 = 1.$

(ii) Closeness Centrality

$$C_C^*(v) = (n-1)\frac{1}{\sum_{j=1} d(v,j)}$$

$$C_C^*(1) = 4 * \frac{1}{7} = 4/7 = 0.571$$

$$C_C^*(2) = 4 * \frac{1}{5} = 4/5 = 0.8$$

$$C_C^*(3) = 4 * \frac{1}{6} = 4/6 = 0.667$$

$$C_C^*(4) = 4 * \frac{1}{4} = 1$$

$$C_C^*(5) = 4 * \frac{1}{6} = 0.667$$

(iii) Betweenness Centrality

A	B	C	D	E
1	2	3	4	5

Betweenness of 1:-

2 → 3	0/1
2 → 4	0/1
2 → 5	0/1
3 → 4	0/1
3 → 5	0/2
4 → 5	0/1

Total = 0

here v = 1

$C_B(v) = 0 * 2$

$C_B^*(v) = \frac{0 * 2}{2 * (4\text{choose}2)} = \frac{0*2}{12} = 0$

Betweenness of 2:-

1 → 3	0/1	3 → 4	0/1
1 → 4	0/1	3 → 5	1/2
1 → 5	0/1	4 → 5	0/1

Total = 1/2

$C_B(2) = 1/2 * 2 = 1$

$C_B^*(2) = \frac{1}{12} = 0.0833$

Betweenness of 3:-

1 → 2	0/1	2 → 4	0/1
1 → 4	0/1	2 → 5	0/1
1 → 5	0/1	4 → 5	0/1

Total = 0

$C_B^*(3) = 0$

Betweenness of 4

1 → 2	0/1	2 → 3	0/1
1 → 3	1/1	2 → 5	0/1
1 → 5	1/1	3 → 5	1/2

Total = 3.5

$C_B(4) = 3.5 * 2 = 7$

$C_B^*(4) = \frac{7}{12} = 0.5833$

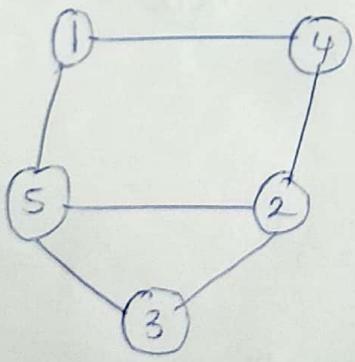
Betweenness of 5:-

1 → 2	0/1	2 → 3	0/1
1 → 3	0/1	2 → 4	0/1
1 → 4	0/1	3 → 4	0/1

Total = 0

$C_B^*(5) = 0$

b)



Degree centrality:-

$C_D^*(1) = 2/4 = 0.5$

$C_D^*(2) = 3/4 = 0.75$

$C_D^*(3) = 2/4 = 0.5$

$C_D^*(4) = 2/4 = 0.5$

$C_D^*(5) = 3/4 = 0.75$

(ii) Closeness centrality

$$C_C^*(v) = (n-1) \cdot \frac{1}{\sum_j d(v,j)}$$

$C_C^*(1) = \frac{4}{6} = 0.667$

$C_C^*(2) = \frac{4}{5} = 0.8$

$C_C^*(3) = \frac{4}{6} = 0.667$

$C_C^*(4) = \frac{4}{6} = 0.667$

$C_C^*(5) = 0.8 \Rightarrow 4/5$

(iii) Betweenness centrality

1 2 3 4 5

Betweenness of 1:-

2 → 3	0/1	3 → 4	0/1
2 → 4	0/1	3 → 5	0/1
2 → 5	0/1	4 → 5	1/2

Total = 0.5

$C_B(1) = 2 * 0.5 = 1$

$C_B^*(1) = \frac{1}{12} = 0.0833$

Betweenness of 2:-

1 → 3	0/1	3 → 4	1/1
1 → 4	0/1	3 → 5	0/1
1 → 5	0/1	4 → 5	1/2

Total = 1.5

$C_B(2) = 3$

$C_B^*(2) = \frac{3}{12} = 0.25$

Betweenness of 3:-

1 → 2	0/2	2 → 4	0/1
1 → 4	0/1	2 → 5	0/1
1 → 5	0/1	4 → 5	0/2

Total = 0

$C_B^*(3) = 0$

Betweenness of 4 :-

1 → 2	1/2	2 → 3	0/1	Total = 0.5
1 → 3	0/1	2 → 5	0/1	CB(4) = 1
1 → 5	0/1	3 → 5	0/1	

$CB^*(4) = 1/12 = 0.0833$

Betweenness of 5 :-

1 → 2	1/2	2 → 3	0/1	Total = 1.5
1 → 3	1/1	2 → 4	0/1	CB(5) = 3
1 → 4	0/1	3 → 4	0/1	

$CB^*(5) = 3/12 = 0.25$